An adaptive method of estimation and outlier detection in regression applicable for small to moderate sample sizes

ARTICLE · JANUARY 2000

DOI: 10.7151/dmps.1002

1 AUTHOR:

Brenton Clarke
Murdoch University
44 PUBLICATIONS **253** CITATIONS

SEE PROFILE

AN ADAPTIVE METHOD OF ESTIMATION AND OUTLIER DETECTION IN REGRESSION APPLICABLE FOR SMALL TO MODERATE SAMPLE SIZES

Brenton R. Clarke

Mathematics and Statistics, Division of Science and Engineering
Murdoch University
Murdoch, W.A., 6150, Australia

Abstract

In small to moderate sample sizes it is important to make use of all the data when there are no outliers, for reasons of efficiency. It is equally important to guard against the possibility that there may be single or multiple outliers which can have disastrous effects on normal theory least squares estimation and inference. The purpose of this paper is to describe and illustrate the use of an adaptive regression estimation algorithm which can be used to highlight outliers, either single or multiple of varying number. The outliers can include 'bad' leverage points. Illustration is given of how 'good' leverage points are retained and 'bad' leverage points discarded. The adaptive regression estimator generalizes its high breakdown point adaptive location estimator counterpart and thus is expected to have high efficiency at the normal model. Simulations confirm this. On the other hand, examples demonstrate that the regression algorithm given highlights outliers and 'potential' outliers for closer scrutiny.

The algorithm is computer intensive for the reason that it is a global algorithm which is designed to highlight outliers automatically. This also obviates the problem of searching out ``local minima'' encountered by some algorithms designed as fast search methods. Instead the objective here is to assess all observations and subsets of observations with the intention of culling all outliers which can range up to as much as approximately half the data. It is assumed that the distributional form of the data less outliers is approximately normal. If this distributional assumption fails, plots can be used to indicate such failure, and, transformations may be ;required before potential outliers are deemed as outliers. A well known set of data illustrates this point.

Keywords: outlier; least median of squares regression; least trimmed squares; trimmed likelihood; adaptive estimation; leverage.

1991 Mathematics Subject Classification: 62T05.

References

[1] A.C. Atkinson, *Two graphical displays for outlying and influential observations in regression*, Biometrika **68** (1981), 13-20.

[2] A.C. Atkinson, *Masking unmasked*, Biometrika **73** (1986a), 533-41.

[3] A.C. Atkinson, *Comment : Aspects of diagnostic regression analysis*, Statistical Science **1** (1986b), 397-401.

[4] A.C. Atkinson, *Fast very robust methods for the detection of multiple outliers*, Journal of the American Statistical Association **89** (1994), 1329-1339.

[5] V. Barnett and T. Lewis, Outliers in Statistical Data, 3rd ed., New York, Wiley, 1994.

[6] T. Bednarski and B.R. Clarke, *Trimmed likelihood estimation of location and scale of the normal distribution*, Australian Journal of Statistics **35**(1993), 141-153.

[7] M.D. Brown, J. Durbin and J.M. Evans, *Techniques for testing the constancy of regression relationships over time*, Journal of the Royal Statistical Society, Series B **37** (1975), 149-192.

[8] K.A. Brownlee, Statistical Theory and Methodology in Science and Engineering, 2nd ed., New York, Wiley, 1965.

[9] R.W. Butler, *Nonparametric interval and point prediction using data trimmed by a Grubbs-type outlier rule*, Annals of Statistics **10** (1982), 197-204.

[10] R.L. Chambers and C.R. Heathcote, *On the estimation of slope and the identification of outliers in linear regression*, Biometrika **68** (1981), 21-33.

[11] B.R. Clarke, *Empirical evidence for adaptive confidence intervals and identification of outliers using methods of trimming*, Australian Journal of Statistics **36** (1994), 45-58.

[12] R.D. Cook and S. Weisberg, Residuals and Influence in Regression, New York and London, Chapman and Hall 1982.

[13] P.L. Davies, *The asymptotics of S-estimators in the Linear Regression Model*, Annals of Statistics **18** (1990), 1651-1675.

[14] P.L. Davies and U. Gather, *The identification of multiple outliers (with discussion)*, Journal of the American Statistical Association **88** (1993), 782-801.

[15] N.R. Draper and H. Smith, Applied Regression Analysis, New York, Wiley, 1966.

[16] W. Fung, *Unmasking outliers and leverage points : A confirmation*, Journal of the American Statistical Association **88** (1993), 515-519.

[17] A.S. Hadi and J.S. Simonoff, *Procedures for the identification of multiple outliers in linear models*, Journal of the American Statistical Association**88** (1993), 1264-1272.

[18] F.R. Hampel, E.M. Ronchetti, P.J. Rousseeuw and W.J. Stahel, Robust Statistics, the Approach Based on Influence Functions, New York, Wiley, 1986.

[19] D.M. Hawkins, D. Bradu and G.V. Kass, *Location of several outliers in multiple-regression data using elemental sets*, Technometrics **26** (1984), 197-208.

[20] T.P. Hettmansperger and S.J. Sheather, *A cautionary note on the method of least median squares*, American Statistician **46** (1992), 79-83.

[21] L.A. Jaeckel, *Some flexible estimates of location*, Annals of Mathematical Statistics **42** (1971), 1540-1552.

[22] F. Kianifard and W.H. Swallow, *Using recursive residuals, calculated on adaptively-ordered observations, to identify outliers in linear regression*, Biometrics **45** (1989), 571-585.

[23] F. Kianifard and W.H. Swallow, *A Monte Carlo comparison of five procedures for identifying outliers in linear regression*, Communications in Statistics, Part A-Theory and Methods **19** (1990), 1913-1938.

[24] M.G Marasinghe, *A multistage procedure for detecting several outliers in linear regression*, Technometrics **27** (1985), 395-399.

[25] P.J. Rousseeuw, *Least median of squares regression*, Journal of the American Statistical Association **79** (1984), 871-880.

[26] P.J. Rousseeuw and A.M. Leroy, Robust Regression and Outlier Detection, New York, Wiley, 1987.

[27] P.J. Rousseeuw and B.C. van Zomeren, *Unmasking multivariate outliers and leverage points*, Journal of the American Statistical Association **85**(1990), 633-651.

[28] P.J. Rousseeuw and V.J. Yohai, *Robust regression by means of S-estimators*, in: Robust and Nonlinear Time Series Analysis, eds., J. Franke, W. Härdle and R.D. Martin, (Lecture Notes in Statistics), New York, Springer-Verlag, (1984), 256-272.

[29] D. Ruppert, *Computing S-estimators for regression and multivariate location/dispersion*, Journal of Computational and Graphical Statistics **1**(1992), 253-270.

[30] T.P. Ryan, *Comment on Hadi and Simonoff*, Letters to the Editor, Journal of the American Statistical Association **90** (1995), 811.

[31] G. Simpson, D. Ruppert and R.J. Carroll, *On one-step GM estimates and stability of inferences in linear regression*, Journal of the American Statistical Association **87** (1992), 439-450.

[32] W.H. Swallow and F. Kianifard, *Using robust scale estimates in detecting multiple outliers in linear regression*, Biometrics **52** (1996), 545-556.

[33] J.W Tukey and D.H. McLaughlin, *Less vulnerable confidence and significance procedures for location based on a single sample: Trimming/Winsorization* 1, Sankhya **25** (A) (1963), 331-352.

[34] W.N. Venables and B.D. Ripley, Modern Applied Statistics with S-Plus, New York, Springer-Verlag, 1994.

[35] D.L. Woodruff and D.M. Rocke, *Computable robust estimation of multivariate location and shape in high dimension using compound estimators*, Journal of the American Statistical Association **89** (1994), 888-896.

[36] V.J. Yohai, *High breakdown point and high-efficiency robust estimates for regression*, Annals of Statistics **15** (1987), 642-656.

[37] V.J. Yohai and R.H. Zamar, *High breakdown-point estimates of regression by means of the minimization of an efficient scale*, Journal of the American Statistical Association **83** (1988), 406-413.

Received 15 November 1998